BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: 11039334 A

(43) Date of publication of application: 12.02.99

(51) Int. Cl G06F 17/30
G06F 17/27
G06F 17/21

(21) Application number: 09197015

(22) Date of filing: 23.07.97

(71) Applicant: CANON INC

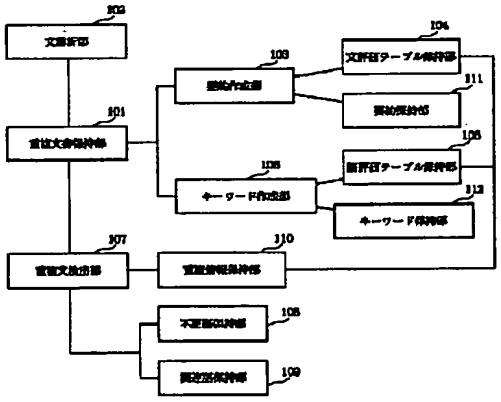
(72) Inventor: SHIBATA SHOGO
MACHIDA NORIKO
ITO SHIRO
UEDA TAKANARI
IKEDA YUJI

(54) DOCUMENT PROCESSOR, ITS METHOD AND STORAGE MEDIUM STORING ITS PROGRAM

(57) Abstract:

PROBLEM TO BE SOLVED: To score information from plural overlapping documents and to create a summary, a keyword, etc., taking objective significance into consideration by extracting information from each document based on an evaluation value of information in plural documents.

SOLUTION: A sentence analyzing part 102 analyzes each sentence in a document held by an overlapping document holding part 101, and an overlapping sentence detecting part 107 detects overlapping in each sentence from an overlapping document. In such cases, an unnecessary word holding part 108 holds a word to be ignored, and a related word holding part 109 holds a word that has strong relation with a certain word. An overlapping information holding part 110 holds overlapping information that is detected by the part 107. A summary creating part 103 creates a summary of a document held by the part 101. A sentence evaluation table that is used in such cases is held by a sentence evaluation table holding part 104. A summary holding part 111 holds a sentence that is created by the part 103.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－39334

(43)公開日　平成11年(1999)２月12日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３２０Ａ
17/27		15/20	５５０Ｆ
17/21			５９０Ｅ

審査請求　未請求　請求項の数15　ＯＬ　（全８頁）

(21)出願番号　特願平9－197015

(22)出願日　平成９年(1997)７月23日

(71)出願人　000001007
キヤノン株式会社
東京都大田区下丸子３丁目30番２号

(72)発明者　柴田　昇吾
東京都大田区下丸子３丁目30番２号キヤノン株式会社内

(72)発明者　町田　紀子
東京都大田区下丸子３丁目30番２号キヤノン株式会社内

(72)発明者　伊藤　史朗
東京都大田区下丸子３丁目30番２号キヤノン株式会社内

(74)代理人　弁理士　丸島　儀一

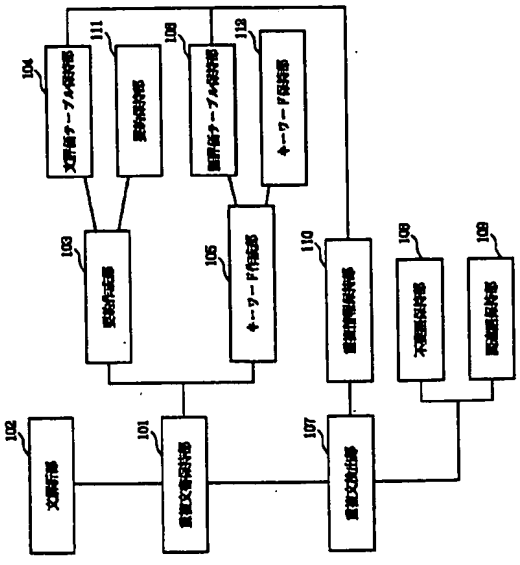
最終頁に続く

(54)【発明の名称】　文書処理装置及び方法、及びそのプログラムを記憶した記憶媒体

(57)【要約】

【課題】　重複のある複数の文書に対し、適切な要約やキーワードを作成する。

【解決手段】　内容の重複する複数の文書中の各文を解析する文解析部110と、解析された複数の文書中から重複部分のある複数の文を検出する重複文検出部106と、検出された重複部分のある複数の文を、当該重複部分に基づいて1つの文に統合し、統合された文に基づいて、前記複数の文書を統合する文統合部109とを備える。



1

【特許請求の範囲】

【請求項１】　複数の文書中の各文を解析する文解析手段と、
該文解析手段により解析された複数の文書中から重複部分を検出する重複部分検出手段と、
該重複部分検出手段により検出された重複部分に基づいて、前記複数の文書の各文書中の情報に評価値を与える評価手段と、
前記評価値に基づいて前記各文書から情報を抽出する抽出手段とを備えることを特徴とする文書処理装置。

【請求項２】　前記情報は文であり、前記抽出手段は、各文書から抽出した文を当該文書の要約として出力することを特徴とする請求項1記載の文書処理装置。

【請求項３】　前記重複部分検出手段が、
前記複数の文書の各文書における形態素毎の出現回数を計数する計数手段と、
共通に出現する形態素を持つ文の組に対し、当該形態素の各文書における出現回数に基づく得点を付与する得点付与手段と、
前記文の組に付与された得点に基づいて、当該文の組に重複部分があるか否かを判定する重複判定手段とを有することを特徴とする請求項２記載の文書処理装置。

【請求項４】　前記情報は単語であり、前記抽出手段は、各文書から抽出した単語を当該文書のキーワードとして出力することを特徴とする請求項1記載の文書処理装置。

【請求項５】　前記各文書単独の内容に基づいて、当該各文書中の情報に第2の評価値を与える第2評価手段を備え、
前記抽出手段が、前記該評価及び前記第2の評価値に基づいて情報を抽出することを特徴とする請求項１記載の文書処理装置。

【請求項６】　種々の内容の文で使用される語を不要語として記憶する不要語記憶手段を備え、前記重複部分検出手段が、当該不要語記憶手段に記憶された不要語以外の重複部分を検出することを特徴とする請求項1記載の文書処理装置。

【請求項７】　ある単語と関連する単語とを記憶する関連語記憶手段を備え、前記重複部分検出手段が、当該関連語記憶手段を参照して、ある単語と関連する単語とを重複部分として検出することを特徴とする請求項1記載の文書処理装置。

【請求項８】　複数の文書中の各文を解析する文解析工程と、
該文解析工程で解析された複数の文書中から重複部分を検出する重複部分検出工程と、
該重複部分検出工程で検出された重複部分に基づいて、前記複数の文書の各文書中の情報に評価値を与える評価工程と、
前記評価値に基づいて前記各文書から情報を抽出する抽出工程とを備えることを特徴とする文書処理方法。

2

【請求項９】　前記情報は文であり、前記抽出工程では、各文書から抽出した文を当該文書の要約として出力することを特徴とする請求項８記載の文書処理方法。

【請求項１０】　前記重複部分検出工程が、
前記複数の文書の各文書における形態素毎の出現回数を計数する計数工程と、
共通に出現する形態素を持つ文の組に対し、当該形態素の各文書における出現回数に基づく得点を付与する得点付与工程と、
前記文の組に付与された得点に基づいて、当該文の組に重複部分があるか否かを判定する重複判定工程とを有することを特徴とする請求項９記載の文書処理方法。

【請求項１１】　前記情報は単語であり、前記抽出工程では、各文書から抽出した単語を当該文書のキーワードとして出力することを特徴とする請求項８記載の文書処理方法。

【請求項１２】　前記各文書単独の内容に基づいて、当該各文書中の情報に第2の評価値を与える第2評価工程を備え、
前記抽出工程では、前記該評価及び前記第2の評価値に基づいて情報を抽出することを特徴とする請求項８記載の文書処理方法。

【請求項１３】　種々の内容の文で使用される語を不要語として記憶する不要語記憶工程を備え、前記重複部分検出工程が、当該不要語記憶工程に記憶された不要語以外の重複部分を検出することを特徴とする請求項８記載の文書処理方法。

【請求項１４】　ある単語と関連する単語とを記憶する関連語記憶工程を備え、前記重複部分検出工程が、当該関連語記憶工程を参照して、ある単語と関連する単語とを重複部分として検出することを特徴とする請求項８記載の文書処理方法。

【請求項１５】　請求項８乃至１４に記載の文書処理方法をコンピュータに実行させるための文書処理プログラムを記憶したことを特徴とする記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、文書処理に関し、特に、重複した内容を有する複数の文書について、文書の要約を作成したり、キーワードを付与したりする文書処理装置及びその方法、及びそのプログラムを記憶する記憶媒体に関するものである。

【０００２】

【従来の技術】今日では、記憶媒体の大容量化・低価格化、ワードプロセッサの普及等によって電子化された文書の量が増大している。さらに、ネットワークの整備が進み、電子メール・電子ニュース等のメディアによってユーザのもとに届く電子化文書の量も増えている。このため、ユーザが処理できる量を越えた文書が入ってくる

ようになるという、いわゆる「情報洪水」が問題になってきている。

【０００３】この問題への対応策として、各文書の量を減らすために文書の内容を要約する「文書要約」技術が用いられるようになってきた。「文書要約」では、例えば文書中の各文に重要度にしたがってスコアを付け、スコアの高いものから文を選択することにより、定められた比率で要約文書を生成する。スコア付けは、段落の先頭にあるか中間にあるかといった文の出現位置や、文章中の重要語を含んでいるかなどを判断基準として行なわれている。

【０００４】また、大量の文書から必要な文書を得るには、各文書にキーワードを付与しておき、キーワード検索により文書を選び出すことも行われている。このためのキーワードを生成する技術として、文書中の各語に重要度にしたがってスコアを付け、スコアの高い語をキーワードとして選択することも行われている。スコア付けは、段落の先頭にあるか中間にあるかといった文の出現位置や、文章中で繰り返し用いられているかなどを判断基準として行なわれている。

【０００５】

【発明が解決しようとする課題】しかしながら、上記従来の要約及びキーワードの作成方法では、一文書内の情報だけをもとにしたスコア付けを行なっていたので、その文書の書き手が重要だと判断した情報のスコアは高くなるものの、必ずしも一般の重要度を反映しているとは言えなかった。

【０００６】また、インターネットの整備等により、情報発信が容易になるとともに、同じ事柄について様々な情報が提供されるようになってきた。例えば、コンピュータの新製品が発表されると、発売したメーカ、新聞等の報道機関、消費者などから、異なった立場で製品についての情報が発信されている。これら同じ事柄に関する情報が個別に要約されると、他の情報との差がわからなくなったり、必要な情報を読み落としてしまうことになりかねないという問題があった。

【０００７】本発明は、上述した従来の課題を解決するためになされたものであり、重複した複数の文書から情報のスコア付けを行ない、客観的な重要度を考慮して要約やキーワードなどを作成することを目的とする。

【０００８】

【課題を解決するための手段】上述した目的を達成するための一手段として、本発明によれば、文書処理装置に、複数の文書中の各文を解析する文解析手段と、該文解析手段により解析された複数の文書中から重複部分を検出する重複部分検出手段と、該重複部分検出手段により検出された重複部分に基づいて、前記複数の文書の各文書中の情報に評価値を与える評価手段と、前記評価値に基づいて前記各文書から情報を抽出する抽出手段とを備える。

【０００９】また、本発明の他の態様によれば、文書処理方法に、複数の文書中の各文を解析する文解析工程と、該文解析工程で解析された複数の文書中から重複部分を検出する重複部分検出工程と、該重複部分検出工程で検出された重複部分に基づいて、前記複数の文書の各文書中の情報に評価値を与える評価工程と、前記評価値に基づいて前記各文書から情報を抽出する抽出工程とを備える。

【００１０】更に、本発明の他の態様によれば、記憶媒体に、前記文書処理方法をコンピュータに実行させるための文書処理プログラムを記憶する。

【００１１】

【発明の実施の形態】

（第１の実施形態）以下、図面を用いて本発明の１実施形態を詳細に説明する。

【００１２】図1は、本発明の実施形態に係る文書処理装置の構成を示すブロック図である。

【００１３】同図において、101は入力された重複文書が保存されている重複文書保持部、102は重複文書保持部101に保持されている文書中の各文を解析する文解析部、103は重複文書保持部101に保持されている文書の要約を作成する要約作成部、104は要約作成部103で使用する文評価のテーブルを保持する文評価テーブル保持部、105は重複文書保持部101に保持されている文書のキーワードを作成するキーワード作成部、106はキーワード作成部105で使用する語評価のテーブルを保持する語評価テーブル保持部、107は重複文書から文毎の重複を検出する重複文検出部、108は重複文を検出する際に無視すべき単語を保持する不要語保持部、109はある単語と関連が強い単語を保持する関連語保持部、110は重複文検出部107で検出した重複情報を保持する重複情報保持部、111は要約文生成部103で生成した文を保持する要約保持部、112はキーワード作成部105で作成した結果を保持するキーワード保持部である。

【００１４】図2は上述の文書処理装置のハードウェア構成を示す図である。

【００１５】同図において、201は、図3～5に示す制御手順に対応するプログラムを記憶する制御メモリである。これはROMであつてもよいし、RAMであつてもよい。202はメモリで、文評価テーブル保持部104と語評価テーブル保持部106と重複情報保持部110とを実現し、上記プログラムの動作に必要な記憶領域とを提供する。203は制御メモリ201に記憶されているプログラムにしたがって処理を行なうCPUである。

【００１６】204はディスク装置であり、重複文書保持部101と不要語保持部108と関連語保持部109とを実現する。205はマウス・キーボード等の入力部である。206は出力部である。これはCRT、液晶ディスプレイ、プリンタ等、どのような装置であってもよい。207は各構成要素を接続するためのバスである。

【００１７】図3は、図1に示した装置における動作の処理手順を示すフローチャートである。図4～図5は、図3の処理手順の詳細を示すフローチャートである。また、図6は本発明の実施形態の具体例であり、図7～図10は文の評価処理を説明するための具体例、図11、図12はキーワードのための語の評価処理を説明するための具体例である。

【００１８】図3～図12を参照しながら、本実施の形態の動作を説明する。本実施の形態では入力文書として、図6に示すような、内容が重複している「文書1」と「文書2」とを想定する。内容が重複している文書は、一般に行なわれているどのような方法で得ても良いものとする。例えば、人間がデータベース等に条件を指定して検索して得たり、興味ある記事をスクラップする等して得ることができる。

【００１９】まず、ステップS301では、重複文書保持部101にある全ての文書について、従来の手法により要約とキーワードの作成を行なう。これらの処理は、ここで作成した要約とキーワードそのものを用いるためではなく、作成の際に文書中の各文や各語を評価するための評価テーブルの作成が目的である。

【００２０】要約を作成するためには、一般に、以下に示す基準で文の評価を行なう。

・文書の冒頭や段落の冒頭の文には高い得点を与える
・例示の段落や文については不要なので減点する
・文書中でよく使われる重要語を含んでいる文に加点する
・重要な内容を表わす語（表現）を含んでいる文に加点する

このようにして評価を行なった結果が文評価テーブル保持部104に格納される。文評価テーブルの1例を図7に示す。文書1の評価結果を701に、文書2の評価結果を702に示す。この表で得点が高いほど重要な文で、単独の要約を得る場合には、高い得点の文を任意の数だけ取り出せばよい。

【００２１】キーワードを作成するためには、一般に、以下に示す基準で語の評価を行なう。

・文書の冒頭や段落の冒頭に出現する語には高い得点を与える
・例示の段落や文に含まれている語については不要なので減点する
・文書中に繰り返し用いられている語には高い得点を与える
・文の主語や目的語に用いられている語には高い得点を与える
・どんな文にも頻繁に用いられる語についてはキーワードから排除する

このようにして評価を行なった結果が語評価テーブル保持部106に格納される。語評価テーブル106の1例を図11に示す。文書1の評価結果を1101に、文書2の評価結果を1102に示す。

【００２２】次に、ステップS302以降で、重複文書同士の文の重複を検出する。同じ文が複数文書で重複しているということは、複数の情報源から「書くべきこと」と判断されたのであるから、重要度が高いと見なしてよいと考える。

【００２３】まず、ステップS302で、重複文書保持部101にある全ての文書について形態素の出現回数をカウントする。各文書の形態素の出現頻度を調べる理由は、例えば、文書1と文書2でそれぞれ1回しか用いられない単語があれば、その単語を含む文同士が一致する可能性が高いので、文書中における出現回数の少ない単語から比較をするのが有効なためである。

【００２４】図4にステップS302の詳細を示す。まず、ステップS401において、文書中の全ての文を解析する。文解析の手法については、一般に行なわれるどのような手法を用いてもよい。解析を行なった結果、文中の形態素が得られる。

【００２５】ステップS402で、未処理の形態素があるかをチェックし、未処理の形態素がなくなれば、処理を終了する。未処理の形態素があれば、その１つについて以下の処理を行なう。

【００２６】次に、ステップS403で、形態素が不要語保持部108に登録されている不要語ではないかをチェックする。不要語とは、「する」「なる」「こと」のように、日本語であればどんな文書にも多用される形態素である。不要語の例を図8に示す。このような形態素を残しておくと、本来なら関係のない文同士でも形態素の合致が起こる危険性があるので、ここで排除する。

【００２７】不要語でなければ、ステップS404で、関連語保持部109に登録されている語かを調べ、登録されていれば、ステップS405で関連語を追加する。この処理は、形態素の一致を調べる際に、同じ概念であるのに異なった表記を用いる形態素を合致させることを目的とする。例えば、ある文書では「アメリカ合衆国」と書かれているのに他の文書では「USA」や「米国」と書かれていても一致がとれることになる。

【００２８】そして、ステップS406で、処理対象の形態素が既出パタンかを調べ、新たなパタンであれば、ステップS408で、形態素リストに登録し、対応するカウント値を１とする。既出パタンであれば、ステップS407で、対応するカウント値を＋１する。

【００２９】以上のように、形態素ごとにカウンタを持たせ文解析結果保持部104に保存する処理を、全ての形態素について行ない、1文書中の頻度が求められる。「形態素のカウント」の結果の一例を図6に示す。

【００３０】ステップS303の「文の重複検出」処理は、重複があったと判定された文書同士で、具体的にどの文とどの文とが対応しているかを検出するものである。この処理の前提として、重複文書である文書1と文書2のい

ずれについてもステップS302の「形態素のカウント」処理が行なわれているものとする。

【００３１】図5にステップS303の詳細を示す。頻度を調べたすべての形態素の一つ一つについて処理を行なうために、まず、ステップS501において、未処理の形態素があるかを調べ、あれば、ステップS502で、その1つに着目する。着目した形態素をxとすると、ステップS503で、重複している文書2がこの形態素を含むかをチェックする。文書2に含まれていなければ、この形態素は文の重複検出には使用できないということになるので、次の形態素に着目する。文書2に含まれていれば、ステップS504で、文書1での形態素xの出現回数（N(回)とする）、文書2での出現回数（M(回)とする）を求める。

【００３２】文書1の文にA、B、C、…という番号をふり、文書2の文に1、2、3、…という番号をふると、例えば、形態素xが文書1のBとEの2箇所に用いられ、文書2では1と3と5の3箇所に用いられた場合、(B，1)，(B，3)，(B，5)，(E，1)，(E，3)，(E,5)という6通りの組み合わせが考えられる。そこで、ステップS505で、これらの組み合わせに均等に得点16（＝100／6）を加える。

【００３３】文書1の全ての形態素について、以上の処理を行なうと、ステップS506において、総得点がT以上であれば文の重複があったとしてステップS508で登録し、T以下であれば文の重複がなかったとしてステップS507でその旨を登録する。この得点の一例を図8に示す。この例では、T＝200であり、(A，2)，(C，3)，(D，4)，(E，1)が重複している文だと見なされたことになる。

【００３４】ステップS304で、重複文だと見なされた文の得点を加える。図10に示すように、この例では、図7の得点に図9の得点の10分の1を加える。例えば図9の(A，2)が200ポイントであるから、Aの文に20ポイントを加えて合計が94ポイント、2の文はもとが0ポイントなので加えた20ポイントが合計となる。

【００３５】次に、この合計から評価の高い文を重要文として選択する。図10では、902の4番目の文が122ポイントで最高点、1番目の文が118でそれに次いでいる。これらの文から重要文をいくつ取り出すかはあらかじめ決めておく。この例では、文字数の少ない方の文書の30%程度の大意を抽出したいので、2文を取り出す。3文を取り出すと30%を超えてしまうのである。結果として、文書2の文(4)と文(1)が選択された。

【００３６】そして、ステップS306で文(4)と文(1)を適正な順序に並べ換える。これらはどちらも同一の文書2から取り出したものなので、文書2の順序通り、「文(1)，文(4)」と並べる。もし、異なった文書からの文を並べる場合は、例えば文書1の文(c)と、文書2の文（1）と文（4）を並べる場合には、多数決で文書2の並べ方を重視して、文書1の文(c)と重複する文(3)の位置に文(c)を挿入し、「文(1)、文(c)、文(4)」と並べる。

【００３７】ステップS307で並べた文を文字列として連結し、要約を作成し、要約保持部111に格納する。

【００３８】次にステップS308でキーワードについての処理を行なう。図12に示すように、ステップS301でのキーワード作成時の評価に、ステップS503で重複があった場合の加点を行なう。この例では、重複があった場合に一律20ポイントを加えている。この結果を合計した得点で高い順に並べ換える。

【００３９】そして、ステップS309で、あらかじめ指定しておいた個数であるN個を得点の高い順に取り出すと、例えば、N＝3とすると、80ポイントの「電子メール」、75ポイントの「電話」、60ポイントの「C社」がキーワードとなる。なお、67ポイントの「携帯電話」と64ポイントの「電子メール」は、それぞれ、「電話」「電子メール」と重複してといるのでキーワードとは採用されない。

【００４０】以上のようにして、重複文書の要約とキーワードを作成する。

【００４１】（他の実施形態）上記実施形態では、要約とキーワードを同時に作成していたが、要約とキーワードをそれぞれ単独に作成しても良いものである。

【００４２】上記実施形態では、2つの文書を対象として重複部分を抜き出していたが、3つ以上の文書を対象としても良いものとする。その際、全部の文書に含まれている部分だけを抜き出すか、任意の2以上の数Nを指定して、N文書に含まれている部分を抜き出すかという指定ができるようにしても良いものとする。

【００４３】上記実施形態では、複数の文書から一つの要約を作成していたが、文書ごとの要約を作成しても良いものである。即ち、文の評価を行なった得点を加えた図10で、1001で上位の文を取り出して文書1の要約を作成しても良い。

【００４４】上記実施形態では、複数の文書から一つのキーワードを作成していたが、文書ごとにキーワードを作成しても良いものである。即ち、語の評価を行なった得点を加えた図12で、1201で上位の語を取り出して文書1のキーワードとしても良い。

【００４５】上記実施形態では、文の評価の得点に、文の重複による得点の10分の1の得点を加えていたが、どんな比率で加えても良いものとする。また、加算だけではなく、比率に変換して乗算としても良いものとする。

【００４６】上記実施形態では、キーワード作成時の重複語の得点の加点を、一律20ポイントとしていたが、可変にしても良いものとする。例えば、最初にキーワードを作成する時の得点に応じて変化させたり、重複の一致度に応じて変化させても良いものとする。

【００４７】上記実施形態では、文書1と文書2との文の関係付けは1文対1文で行なっていたが、1文対複数文、あるいは複数文対1文でも良いものとする。これは、例えば、文書1では一つの文であっても、文書2では複数の場所に分けて記述されていることがあるからである。1

文と複数文との重複部分を抜き出す処理については、1文対1文と同様に、1文と複数文との長さを比較して短い方を取り出し重複部分を残していく。

【００４８】上記実施形態では、日本語の文書を対象としていたが、英語、スペイン語等、他の言語の文書を対象とするのも良いものとする。

【００４９】上記実施形態では、文の重複検出の際に、同一の得点（本実施形態では100ポイント）を出現パタンの数で割っていたが、合致の状況に応じて得点を変えても良いものとする。例えば、名詞の合致得点に較べて動詞の合致得点を高くするとか、名詞でも文の主語となっている格要素の合致であれば得点を高くすることが有効である。

【００５０】上記実施形態では、文の重複検出の際に、複合名詞、複合動詞等の複合語について、特に処理を分けていなかったが、複合語については複合語の特性に応じた処理を行なっても良いものとする。例えば、「自然言語処理技術」に対して、「言語処理技術」であれば7割、「処理技術」であれば3割といった得点を与える。

【００５１】また、上記実施形態では、文の重複検出の際に、関連語についても同一の得点を与えていたが、関連の度合いに応じて得点を変えても良いものとする。例えば、「アメリカ合衆国」に対して「USA」は10割であるが、「欧米」であれば5割、「先進7か国」であれば7割というような得点を与える。

【００５２】なお、本発明は、複数の機器から構成されるシステムに適用しても、1つの機器からなる装置に適用してもよい。前述した実施形態の機能を実現するソフトウェアのプログラムコードを記録した記録媒体を、システム或いは装置に供給し、そのシステム或いは装置のコンピュータ（またはCPUやMPU）が記録媒体に格納されたプログラムコードを読み出し実行することによっても、達成されることは言うまでもない。

【００５３】この場合、記録媒体から読み出されたプログラムコード自体が前述した実施形態の機能を実現することになり、そのプログラムコードを記録した記録媒体は本発明を構成することになる。

【００５４】上記プログラムコードを供給するための記録媒体としては、例えば、フロッピーディスク、ハードディスク，光ディスク，光磁気デイスク，CD—ROM，CD—R，磁気テープ，不揮発性のメモリカード，ROMなどを用いることができる。

【００５５】また、コンピュータが読み出したプログラムコードを実行することにより、前述した実施形態の機能が実現される場合だけでなく、そのプログラムコードの指示に基づき、コンピュータ上で稼働しているOSなどが実際の処理の一部または全部を行ない、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【００５６】更に、記録媒体から読み出されたプログラムコードが、コンピュータに挿入された機能拡張ボードやコンピュータに接続された機能拡張ユニットに備わるメモリに書き込まれた後、そのプログラムコードの指示に基づき、その機能拡張ボードや機能拡張ユニットに備わるCPUなどが実際の処理の一部または全部を行ない、その処理によって前述した実施形態の機能が実現される場合も含まれることは言うまでもない。

【００５７】

【発明の効果】以上説明したように、本発明によれば、重複のある文書を対象として、重複に関わる部分を優先して抽出することで、文書の要約やキーワードなどが作成できるという効果がある。

【図面の簡単な説明】

【図１】本発明に係る一実施形態の文書処理装置の基本構成を示すブロック図である。

【図２】本発明の実施形態に係るハードウェア構成を示すブロック図である。

【図３】本発明の実施形態に係る処理手順を示すフローチャートである。

【図４】形態素のカウント処理の処理手順を示すフローチャートである。

【図５】文の重複検出処理の処理手順を示すフローチャートである。

【図６】形態素のカウント結果の例を示す図である。

【図７】文の得点の例を示す図である。

【図８】不要語の例を示す図である。

【図９】文の組の得点の例を示す図である。

【図１０】重複による加算後の文の得点の例を示す図である。

【図１１】語の得点の例を示す図である。

【図１２】重複による加算後の語の得点の例を示す図である。

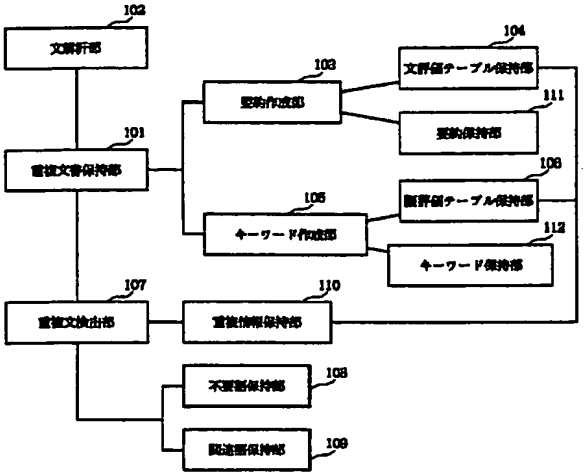
【符号の説明】

１０１　重複文書保持部
１０２　文解析部
１０３　要約作成部
１０４　文評価テーブル保持部
１０５　キーワード作成部
１０６　語評価テーブル保持部
１０７　重複文書検出部
１０８　不要語保持部
１０９　関連語保持部
１１０　重複情報保持部
１１１　要約保持部
１１２　キーワード保持部
２０１　制御メモリ
２０２　メモリ
２０３　ＣＰＵ
２０４　ディスク装置
２０５　入力部

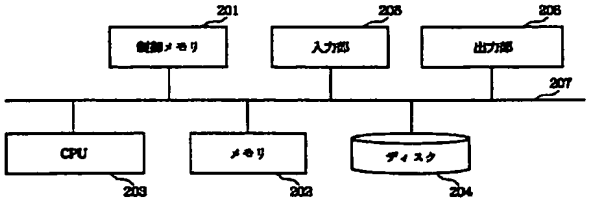
11　　12

206　出力部　　207　バス

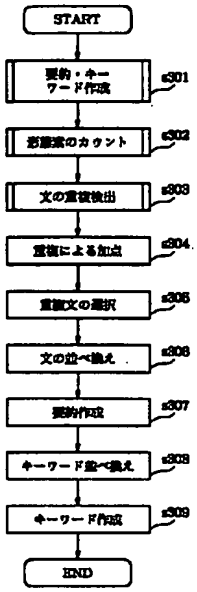
【図1】



【図2】



【図3】



【図7】

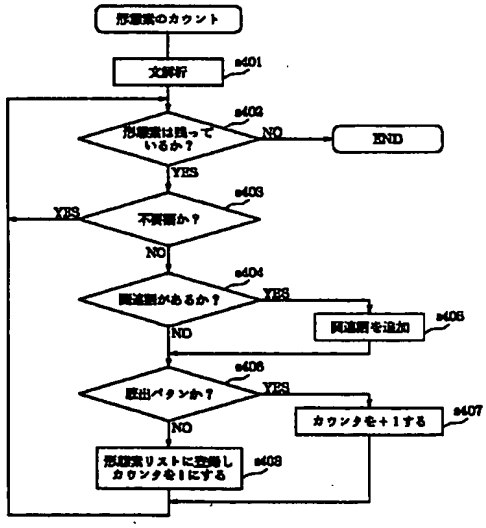
701

	得点
A	74
B	55
C	6
D	0
E	45
F	30

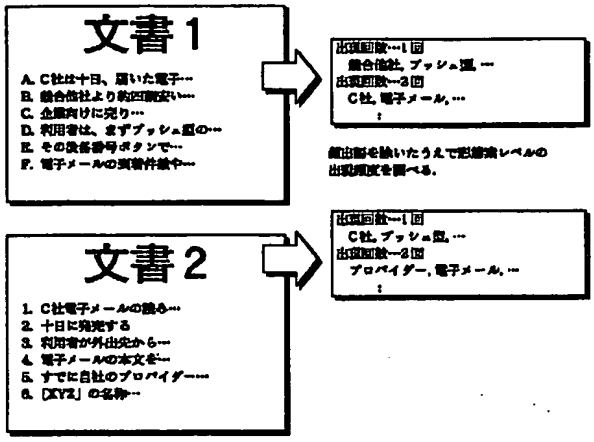
702

	得点
1	88
2	0
3	80
4	50
5	10
6	86

【図4】

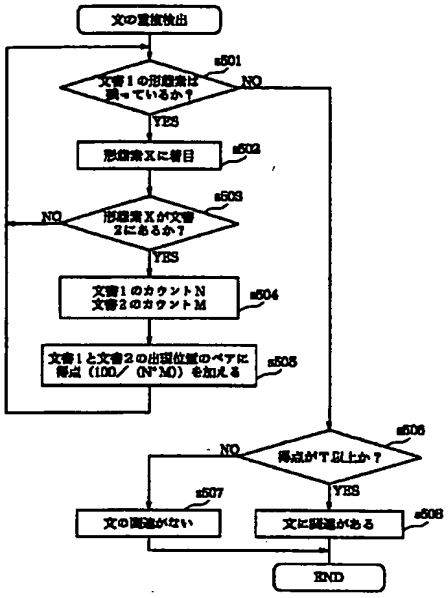


【図6】



重出語を除いたうえで形態素レベルの出現頻度を調べる。

【図５】



【図８】

表記	品詞
こと，事	名詞
なる	動詞
ある	動詞
する	動詞
にほん，日本	名詞
関連	名詞
記事	名詞
番号	名詞
ない	形容詞
：	：

【図９】

	A	B	C	D	E	F
1	66				■	
2	■					
3	116		■			
4				■		
5						
6						

【図１１】

1101

	得点
電子メール	80
電話	55
利用者	50
C社	40
プッシュ型	28
企業向け	20

1102

	得点
プロバイダー	50
携帯電話	47
電子メール	44
プッシュ型	35
外出先	27
C社	24

【図１０】

1001

	得点	重複	合計
A	74	20	94
B	55	0	55
C	5	60	65
D	0	72	72
E	45	33	78
F	30	0	30

1002

	得点	重複	合計
1	85	33	■
2	0	20	20
3	30	65	95
4	60	72	■
5	10	0	10
6	36	0	36

【図１２】

1201

	得点	重複	合計
電子メール	80	20	80
電話	55	20	75
利用者	50	0	50
C社	40	20	60
プッシュ型	28	20	48
企業向け	20	0	20

1202

	得点	重複	合計
プロバイダー	50	0	50
携帯電話	47	20	67
電子メール	44	20	64
プッシュ型	35	20	55
外出先	27	0	27
C社	24	20	44

フロントページの続き

(72)発明者　上田　隆也
東京都大田区下丸子３丁目30番２号キヤノン株式会社内

(72)発明者　池田　裕治
東京都大田区下丸子３丁目30番２号キヤノン株式会社内